# Chapter 96
# Social Network Integration in Document Summarization

**Atefeh Farzindar**
*NLP Technologies Inc., Canada & Université de Montréal, Canada*

## ABSTRACT

*In this chapter, the author presents the new role of summarization in the dynamic network of social media and its importance in semantic analysis of social media and large data. The author introduces how summarization tasks can improve social media retrieval and event detection. The author discusses the challenges in social media data versus traditional documents. The author presents the approaches to social media summarization and methods for update summarization, network activities summarization, event-based summarization, and opinion summarization. The author reviews the existing evaluation metrics for summarization and the efforts on evaluation shared tasks on social data related tracks by ACL, TREC, TAC, and SemEval. In conclusion, the author discusses the importance of this dynamic discipline and great potential of automatic summarization in the coming decade, in the context of changes in mobile technology, cloud computing, and social networking.*

## 1. INTRODUCTION

Automatic summarization of traditional media such as written press and articles has been a popular research domain over the past 25 years. Document summarization is typically performed to save reading time by reducing the amount of information presented to users. Several online news agencies use clustering techniques to categorize news articles and provide pseudo-summaries. In addition, summarizing specific types of documents, such as legal decisions, drew a lot of attention in the research field and the marketing of automatic systems (Farzindar and Lapalme 2004). The purpose of these approaches is to exploit the thematic structure of documents in order to improve coherence and readability of the summary. In recent years, we have been facing new challenges in processing social media data and its integration in document summarization. Texts in social media are extremely noisy, ungrammatical; they do not adhere to conventional rules and they are subject to continuously changing conventions.

DOI: 10.4018/978-1-4666-8614-4.ch096

Over the past few years, online social networking sites (Facebook, Twitter, Youtube, Flickr, MySpace, LinkedIn, Metacafe, Vimeo, etc.) have revolutionized the way we communicate with individuals, groups and communities, and altered everyday practices (Boyd and Ellison, 2007). Nearly one in four people worldwide will use social networks in 2013, according to an eMarketer report (New Media Trend Watch, 2013), "Worldwide Social Network Users: 2013 Forecast and Comparative Estimates." Social media has become a primary source of intelligence because it has become the first response to key events issued by highly dynamic contents generated by 1.73 billion users in 2013. Social media statistics for 2012 has shown that Facebook has grown to more than 800 million active users, adding more than 200 million in a single year. Twitter now has 100 million active users and LinkedIn has over 64 million users in North America alone (Digital Buzz, 2012). Recently, workshops such as Semantic Analysis in Social Media (Farzindar and Inkpen, 2012) and NAACL/HLT workshop on Language Analysis in Social Media (Farzindar et Al. 2013) have been increasingly focusing on the impact of social media on our daily lives, both on a personal and a professional level.

Social media data is the collection of open source information which can be obtained publicly via Blogs and micro-blogs, Internet forums, user-generated FAQs, chat, podcasts, online games, tags, ratings and comments. Social media data has several properties: the nature of conversation is social which are posted in real-time. Geo-locating a group of topically-related conversations is important as it includes emotions, neologisms, credibility/rumors and incentives. The texts are non-structured and are presented in many formats and written by different people in many languages and styles. Also the typography mistakes and chat slang have become increasingly prevalent on social networking sites like Facebook and Twitter. The authors are not professional writers and the pockets of sources in thousands of places on the www.

Monitoring and analyzing this rich and continuous flow of user-generated content can yield unprecedentedly valuable information, which would not have been available from traditional media outlets. Summarization can play a key role in semantic analysis of social media and Social Media Analytics. This has given rise to the emerging discipline of Social Media Analytics, which draws from Social Network Analysis, Machine Learning, Data Mining, Information Retrieval (IR), automatic summarization, and Natural Language Processing (NLP) (Melville et al. 2009).

In the context of analyzing social networks and document summarization, finding powerful methods and algorithms to extract the relevant data in large volumes, various and free formats from multiple sources and languages, is a scientific challenge. Automatic processing and summarization of such data needs to evaluate the appropriate research methods for information extraction, automatic categorization, clustering, indexing data and statistical machine translation.

The sheer volume of social media data and the incredible rate at which new content is created makes manual summarization, or any other meaningful manual analysis, largely infeasible. In many applications the amount of data is too large for effective real-time human evaluation and analysis of the data for a decision maker.

Traditionally, a distinction is made between extractive and abstractive summaries. The former is defined as consisting entirely of content extracted from the input, while the latter contains some content not present in the source (e.g. paraphrased material) (Mani, 2001) and (Mani and Maybury, 1999).

In social media, the real time event search and the need for event detection raise an important issue (Farzindar and Khreich 2013). The purpose of dynamic information retrieval and real time event searches is to effectively execute search strategies on many features, where search queries consider multiple dimensions including their spatial and temporal relationship. In this case, the summari-